Final Project

Pre-requisite

- Understanding of Python
- Understanding of Data Cleaning

Level of Exercise: Beginner

Duration: 4 hours

Project Details:

Objective:

In this exericise, you will be working an Open Dataset dataset coming from Airbnb. Some of the tasks include

- Data Cleaning.
- Data Transformation

Overview of Airbnb Data:

People's main criteria when visiting new places are reasonable accommodation and food. Airbnb (Air-Bed-Breakfast) is an online marketplace created to meet this need of people by renting out their homes for a short term. They offer this facility at a relatively lower price than hotels. Further people worldwide prefer the homely and economical service offered by them. They offer services across various geographical locations

Dataset Source

You can get the dataset for this assessment using the following link:
https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

This dataset contains information such as the neighborhood offering these services, room type, price,avaliabilty, reviews, service fee, cancellation policy and rules to use the house. This analysis will help airbnb in improving its services.

So all the best for your Data Journey on Airbnb data!!!

```
Start coding or generate with AI.
```

Start coding or generate with AI.

Task 1: Data Loading (Python)

1. Read the csv file and load it into a pandas dataframe.
2. Display the first five rows of your dataframe.
3. Display the data types of the columns.

```
## Import Libraries
import pandas as pd
import numpy as np
```

```
## Read the csv file
url = 'C:\\Users\\athir\\OneDrive\\Documents\\Athira Doc\\Airbnb_Open_Data.csv'
data = pd.read_csv(url)
```

```
C:\Users\athir\AppData\Local\Temp\ipykernel_10604\3862898650.py:3: DtypeWarning: Columns
  data = pd.read_csv(url)
```

```
## Display the first 5 rows
print(f"First five raws is {data.head(5)}")
```

```
First five raws is         id                                                NAME      hos
0  1001254                Clean & quiet apt home by the park  80014485718
1  1002102                           Skylit Midtown Castle  52335172823
2  1002403              THE VILLAGE OF HARLEM....NEW YORK !  78829239556
3  1002755                                             NaN  85098326012
4  1003689  Entire Apt: Spacious Studio/Loft by central park  92037596077

  host_identity_verified host name neighbourhood group neighbourhood  \
0            unconfirmed  Madaline            Brooklyn    Kensington
1               verified     Jenna           Manhattan       Midtown
2                    NaN     Elise           Manhattan        Harlem
3            unconfirmed     Garry            Brooklyn  Clinton Hill
4               verified    Lyndon           Manhattan   East Harlem

        lat      long        country  ... service fee minimum nights  \
0  40.64749 -73.97237  United States  ...       $193            10.0
1  40.75362 -73.98377  United States  ...        $28            30.0
2  40.80902 -73.94190  United States  ...       $124             3.0
3  40.68514 -73.95976  United States  ...        $74            30.0
4  40.79851 -73.94399  United States  ...        $41            10.0

  number of reviews last review  reviews per month review rate number  \
0               9.0  10/19/2021               0.21                4.0
1              45.0   5/21/2022               0.38                4.0
2               0.0         NaN                NaN                5.0
```

```
3             270.0    7/5/2019               4.64                4.0
4               9.0  11/19/2018               0.10                3.0

  calculated host listings count  availability 365  \
0                            6.0             286.0
1                            2.0             228.0
2                            1.0             352.0
3                            1.0             322.0
4                            1.0             289.0

                                        house_rules license
0  Clean up and treat the home the way you'd like...     NaN
1  Pet friendly but please confirm with me if the...     NaN
2  I encourage you to use my kitchen, cooking and...     NaN
3                                                NaN     NaN
4  Please no smoking in the house, porch or on th...     NaN

[5 rows x 26 columns]
```

```
## Display the data types
print(data.dtypes)
```

```
id                                int64
NAME                             object
host id                           int64
host_identity_verified           object
host name                        object
neighbourhood group              object
neighbourhood                    object
lat                             float64
long                            float64
country                          object
country code                     object
instant_bookable                 object
cancellation_policy              object
room type                        object
Construction year               float64
price                            object
service fee                      object
minimum nights                  float64
number of reviews               float64
last review                      object
reviews per month               float64
review rate number              float64
calculated host listings count  float64
availability 365                float64
house_rules                      object
license                          object
dtype: object
```

Task 2a: Data Cleaning

1. Drop some of the unwanted columns. These include `host id`, `id`, `country` and `country code` from the dataset.

Please include the code in the cells below.

```
columns_to_remove=['host id','country','country code']
```

```
data.drop(columns=columns_to_remove,inplace=True)
```

```
print(data.columns)
```

```
Index(['id', 'NAME', 'host_identity_verified', 'host name',
       'neighbourhood group', 'neighbourhood', 'lat', 'long',
       'instant_bookable', 'cancellation_policy', 'room type',
       'Construction year', 'price', 'service fee', 'minimum nights',
       'number of reviews', 'last review', 'reviews per month',
       'review rate number', 'calculated host listings count',
       'availability 365', 'house_rules', 'license'],
      dtype='object')
```

Task 2b: Data Cleaning

- Check for missing values in the dataframe and display the count in ascending order. **If the values are missing, impute the values as per the datatype of the columns.**
- Check whether there are any duplicate values in the dataframe and, if present, remove them.
- Display the total number of records in the dataframe before and after removing the duplicates.

```
## Check for missing values in the dataframe and display the count in ascending order.
print(f"Missing values count {data.isnull().sum().sort_values(ascending=True)}")
```

```
Missing values count id                                0
room type                              0
lat                                    8
long                                   8
neighbourhood                         16
neighbourhood group                   29
cancellation_policy                   76
instant_bookable                     105
number of reviews                    183
Construction year                    214
price                                247
NAME                                 250
service fee                          273
host_identity_verified               289
calculated host listings count       319
review rate number                   326
```

```
host name                         406
minimum nights                    409
availability 365                  448
reviews per month               15879
last review                     15893
house_rules                     52131
license                        102597
dtype: int64
```

```
## Check whether there are any duplicate values in the dataframe and if present remove them.
if data.duplicated().any()==True:
    data.drop_duplicates()
    print("duplicate data is removed")
else:
    print("There is no data to be removed")
```

```
duplicate data is removed
```

```
## Display the total number of records in the dataframe after removing the duplicates.
print(f"Total number of records in python is {len(data)}")
```

```
Total number of records in python is 102599
```

Task 3: Data Transformation

- Rename the column `availability 365` to `days_booked`
- Convert all column names to lowercase and replace the spaces in the column names with an underscore "_".

Please include the code in the cells below.

```
## Rename the column.
data=data.rename(columns={'availability 365':'days_booked'})
```

```
## Convert all column names to lowercase and replace the spaces with an underscore "_"
data.columns=data.columns.str.lower()
print(data.head(10))
```

```
         id                                              name  \
0   1001254                Clean & quiet apt home by the park
1   1002102                             Skylit Midtown Castle
2   1002403               THE VILLAGE OF HARLEM....NEW YORK !
3   1002755                                               NaN
4   1003689  Entire Apt: Spacious Studio/Loft by central park
5   1004098         Large Cozy 1 BR Apartment In Midtown East
6   1004650                                   BlissArtsSpace!
7   1005202                                   BlissArtsSpace!
8   1005754                  Large Furnished Room Near B'way
```

```
9  1006307               Cozy Clean Guest Room - Family Apt

  host_identity_verified host name neighbourhood group       neighbourhood  \
0            unconfirmed  Madaline            Brooklyn          Kensington
1               verified     Jenna           Manhattan             Midtown
2                    NaN     Elise           Manhattan              Harlem
3            unconfirmed     Garry            Brooklyn        Clinton Hill
4               verified    Lyndon           Manhattan         East Harlem
5               verified  Michelle           Manhattan         Murray Hill
6                    NaN   Alberta            Brooklyn  Bedford-Stuyvesant
7            unconfirmed      Emma            Brooklyn  Bedford-Stuyvesant
8               verified    Evelyn           Manhattan      Hell's Kitchen
9            unconfirmed      Carl           Manhattan     Upper West Side

        lat      long instant_bookable cancellation_policy  ... service fee  \
0  40.64749 -73.97237            False              strict  ...       $193
1  40.75362 -73.98377            False           moderate  ...        $28
2  40.80902 -73.94190             True           flexible  ...       $124
3  40.68514 -73.95976             True           moderate  ...        $74
4  40.79851 -73.94399            False           moderate  ...        $41
5  40.74767 -73.97500             True           flexible  ...       $115
6  40.68688 -73.95596            False           moderate  ...        $14
7  40.68688 -73.95596            False           moderate  ...       $212
8  40.76489 -73.98493             True             strict  ...       $204
9  40.80178 -73.96723            False             strict  ...        $58

   minimum nights number of reviews last review  reviews per month  \
0            10.0               9.0  10/19/2021               0.21
1            30.0              45.0   5/21/2022               0.38
2             3.0               0.0         NaN                NaN
3            30.0             270.0    7/5/2019               4.64
4            10.0               9.0  11/19/2018               0.10
5             3.0              74.0   6/22/2019               0.59
6            45.0              49.0   10/5/2017               0.40
7            45.0              49.0   10/5/2017               0.40
8             2.0             430.0   6/24/2019               3.47
9             2.0             118.0   7/21/2017               0.99

   review rate number calculated host listings count  days_booked  \
0                 4.0                             6.0        286.0
1                 4.0                             2.0        228.0
2                 5.0                             1.0        352.0
3                 4.0                             1.0        322.0
4                 3.0                             1.0        289.0
5                 3.0                             1.0        374.0
6                 5.0                             1.0        224.0
7                 5.0                             1.0        219.0
8                 3.0                             1.0        180.0
```

```
data.columns=data.columns.str.replace(' ','_')
```

Task 4: Exploratory Data Analysis

- List the count of various room types avaliable in the dataset.
- Which room type has the most strict cancellation policy?

Please include the code in the cells below.

```
## List the count of various room types avaliable with Airbnb
room_type_counts = data['room_type'].value_counts()
print(room_type_counts)
```

```
Entire home/apt    53701
Private room       46556
Shared room         2226
Hotel room           116
Name: room_type, dtype: int64
```

```
## Which room type adheres to more strict cancellation policy
#print(data['room_type'].head(30),data['cancellation_policy'].head(30))
#for i in data['cancellation_policy']:
    #if i=='strict':
m=data.groupby(['room_type','cancellation_policy']).size().unstack(fill_value=0)
print(m)
v=m['strict'].max()
print(v)
positions = m[m.isin([v])].stack().index.tolist()
print(positions)
print(f"Room Type adheres to more strict cancellation policy is {positions[0][0]}")



#print(m.eq([m['strict'].max()]))
#for i,j in zip()
```

```
cancellation_policy  flexible  moderate  strict
room_type
Entire home/apt         17911     17912   17828
Hotel room                 44        37      35
Private room            15376     15652   15505
Shared room               743       742     738
17828
[('Entire home/apt', 'strict')]
Room Type adheres to more strict cancellation policy is Entire home/apt
```

Start coding or generate with AI.

Start coding or generate with AI.

Start coding or generate with AI.